IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

A Hybrid Linguistic and Knowledge based Analysis Approach for Fake News Detection in Social Media

Ms.Roqia tabassum,Asst.Professor*, Akella viswaja,Rangaraju Varsha***, Gorantla Jahnavi******

*(CSE Department,Sphoorthy Engineering College , Nadergul)
Email : roqia041@gmail.com
**(CSE Department , Sphoorthy Engineering College , Nadergul)
Email: aviswaja@gmail.com
*** (CSE Department , Sphoorthy Engineering College , Nadergul)
Email : varshi.r.rangaraju@gmail.com
**** (CSE Department , Sphoorthy Engineering College , Nadergul)
Email: jahnavireddy216@gmail.com

***Abstract*— The rapid development of different social media and content-sharing platforms has been largely exploited to spread misinformation and fake news that make people believe in harmful stories, which can influence public opinion, and could cause panic and chaos among the population. Thus, fake news detection has become an important research topic, aiming at flagging a specific content as fake or legitimate.**

The fake news detection solutions can be divided into three main categories: content-based, social context-based, and knowledge-based approaches. In this project, we propose a novel fake news detection system that uses linguistic based approaches and inherits their advantages, by employing linguistic features (i.e., title, number of words, reading ease, lexical diversity and sentiment), The proposed system only employs five features, which is less than most of the state-of-the-art approaches.

Keywords— Fake news, false information, deception detection, social media, information manipulation, Network Analysis, Linguistic Cue, Fact-checking, LRC, DTC, GBC, RFC, Semantic Analysis.

I. INTRODUCTION

Fake News contains misleading information that could be checked. This maintains lie about a certain statistic in a country or exaggerated cost of certain services for a country, which may arise unrest for some countries like in Arabic spring [1]. There are organizations, like the House of Commons and the Crosscheck project, trying to deal with issues as confirming authors are accountable. However, their scope is so limited because they depend on human manual detection, in a globe with millions of articles either removed or being published every minute, this cannot be accountable or feasible manually. A solution could be, by the development of a system to provide a credible automated index scoring, or rating for credibility of different publishers, and news context. This paper proposes a methodology to create a model that will detect if an article is authentic or fake based on its words, phrases, sources and titles, by applying supervised machine learning algorithms on an annotated (label) dataset, that are manually classified and guaranteed. Then, feature selection methods are applied to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results. We propose to create the model using different classification algorithms. The product model will test the unseen data, the results will be plotted, and accordingly, the product will be a model that detects and classifies fake articles and can be used and integrated with any system for future use. Early, approaches were mainly based on linguistic-based techniques, which rely on language usage and its analysis to predict deception.

The goal of these approaches is to look for instances of leakage found in the content of a text at different levels (i.e., words, sentences, characters, and documents levels). These approaches implement different methods such as: data representation, deep syntax, sentiment, and semantic analyses]. In data representation methods, each word is considered as a single unit, and individual words are aggregated and analysed to reveal linguistic cues of deception. In deep syntax methods, the sentences are converted into a set of rewritten rules (i.e., parse tree) in order to describe the syntax structure. The semantic analysis determines the truthfulness of authors, which describes the degree of compatibility of personal experience compared to the content derived from a collection of analogous data. Finally, sentiment analysis focuses on the extraction of opinion, which involves examining written texts about people's attitudes, sentiments, and evaluations using analytical techniques.

The proposed implements four different machine learning algorithms namely Random Forest (RF) [2], Logistic Regression (LR), Decision Tree, and XG-Boost. The earlier mentioned learning algorithms are trained and tested using different combinations of the aforementioned features, and the most performing classifier is selected[3].

II. PROBLEM STATEMENT

In this project, an attempt is made for the binary classification of the news items as fake or real by using sk learn passive aggressive classifier. The classifier is trained and tested with a data set comprising of fake as well as real news items.

The Statement of this project detection is to comprising four moduli. The first module deals with training the passive aggressive classifier and predicting its accuracy with a test data set. The second module deals with the generation of the confusion matrix for the test data set. The third module deals with the classification of a fake news item through its confusion matrix.

The fourth module deals with the classification of a real news item through its confusion matrix. The confusion matrix of passive aggressive binary classification is a two-by-two matrix. This project use sk learn passive aggressive classifier for generating the confusion matrix. Sk learn features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the python numerical and scientific libraries NumPy and SciPy.

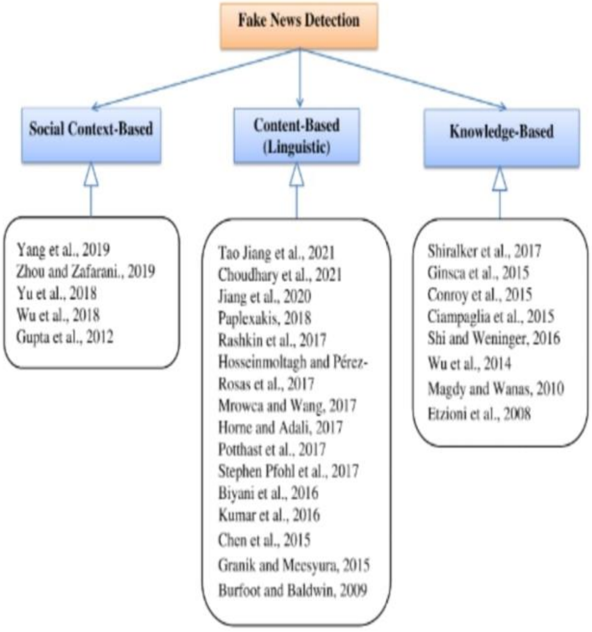
III. LITERATURE SURVEY

Fake news detection approaches can be divided into three categories namely, linguistic-based, social context-based, and knowledge-based. In the figure, some selected approaches from the literature are shown under each category, considered to be the most relevant ones in the last fifteen (15) years.

Here, in literature survey are considering the three types of based analysis which fake news detection will analysis during the process of detecting the news on social media. Through some data sets representation examined the data values to take the review of under solutions.

The three based analysis given as follow shown in the below figure which shows the different variation process of detection are as follows:

(a)Linguistic based analysis: In linguistic based analysis refers to the typical and the accurate examination of natural language. This approach extracts valuable data from the news content, and examines the associated language patterns, meanings and structures of the news. As explained in , linguistic analysis mainly aims to identify the language competence of the news creator by the cognition of language formats and finding out the writing patterns[7].



Data representation, deep syntax, semantic analysis and sentiment analysis are the main used techniques in linguistic analysis. Many studies were conducted to examine the unique linguistic styles in clickbait articles such as in Biyani *et al*. Chen *et al* examined potential methods for the auto- metric detection of clickbait, which aim to find out both textual and non textual clickbait hints among images and users' behaviours.

(b) Knowledge-based analysis: It aims to complement the content– based approaches such as the linguistic ones, by checking the existing body of human knowledge to estimate the likelihood of new statements to be false. The method allows to collect and compare a large number of common and connected statements from different networks like metatags and social network behaviour to compute the probability that the content is fake. According to, knowledge–based analysis and particularly fact checking, aims at using external sources to check the truthfulness of claims in news contents. Magdy and Wanas measured the support for each fact of the document using web search. The measured supports are accumulated to compute the support of the document. According to Ginsca [8] *et al.*

The technique has to take into consideration the different aspects of web information credibility such as: quality, expertise, trustworthiness, and reliability. Etzioni *et al.* proposed an approach for fake news detection based on knowledge analysis, which consists of matching the claims extracted from the web with the analsed news story.

(C) Social Context-based analysis:

The social context-based approaches typically analyse the spreading patterns and the diffusion on social networks to distinguish misleading substance. Yang *et al.* proposed an unsupervised approach for fake news detection on social media. The authors investigated the veracity of news and credibility of users, and utilized a probabilistic graphical model to capture the complete generative spectrum. They evaluated the model on two different datasets (i.e., LIAR and Buzz Feed News), and obtained an accuracy of 75.9% and 67.9% respectively[5].

IV. PROPOSED METHOD

This section presents the methodology used for the classification. Using this model, a tool is implemented for detecting the fake articles. In this method supervised machine learning is used for classifying the dataset. The first step in this classification problem is dataset collection phase, followed by pre-processing, implementing features selection, then perform the training and testing of dataset and finally running the classifiers. Below figure describes the proposed system methodology.

The methodology is based on conducting various experiments on dataset using the algorithms described in the previous section named Random Forest, SVM and Naïve Bayes, majority voting and other classifiers. The experiments are conducted individually on each algorithm, and on combination among them for the purpose of best accuracy precision.

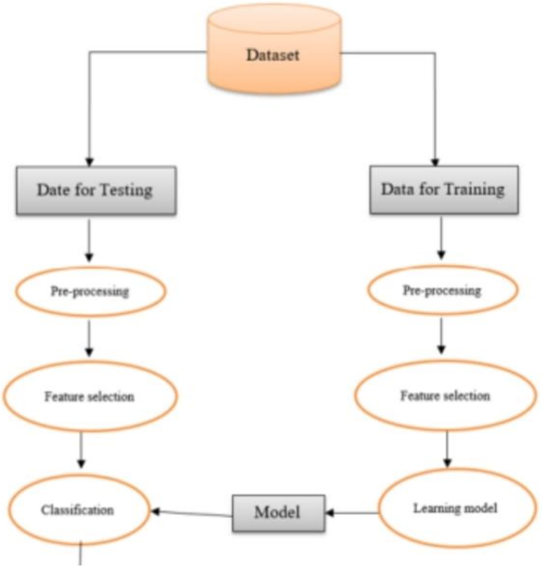


Figure 1. Proposed System Methodology

SYSTEM DESIGN:

In this system detection we have the system methodology where its describes the architecture of entire process of fake news detection on social media as shown in the figure , as the figure shows the algorithms classification where we are using four types of passive aggressive classification. Now, here we need to design the system structure of entire classification.

Now, here we need some structure to design the detection system and also in this system detection we have the system methodology where its describes the to design the system structure of entire classification, architecture of entire process of fake news detection on social media as shown in the figure, as the figure shows the algorithms classification where we are using four types of passive aggressive classification.

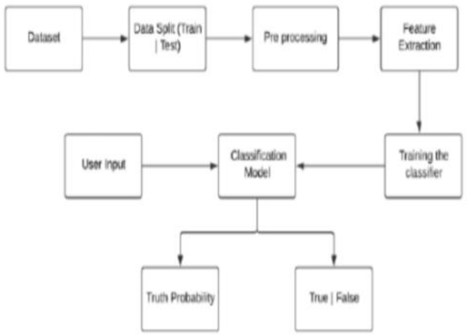


Figure 2. System Design Architecture

In this system architecture, here we have the UML designs where it gives the modeling language used by software developers are of two types are Structural modelling and Behavioral modeling, behavioral are activity , Interaction, use care diagrams and structural are classes, objects, package diagrams.

V. SYSTEM REQUIREMENTS

In hardware there are some requirements used in overview of technology and also procedure software requirements in these Functional requirements.

Hardware requirements:

- Minimum: 2.16 GHZ processor
- Minimum: 4GB RAM
- 64-bit architecture
- Min storage: 500GB

Software requirements:

- 64-bit windows OS
- Python QT designer
- Sklearn

Star UML - Used for creating UML diagrams (DFD and Sequence Diagrams)
Jupyter Notebook - Implementing the python code.

Libraries and Packages Used:

- Pandas
- Seaborn
- Matplotlib
- Sklearn
- NLTK
- Re importing
- String
- Numpy

VI. IMPLEMENTATION

The main goal is to apply a set of classification algorithms to obtain a classification model in order to be used as a scanner for a fake news by details of news detection and embed the model in python application to be used as a discovery for the fake news data. Also, appropriate refactoring have been performed on the Python code to produce an optimized code.

The classification algorithms applied in this model are k-Nearest Neighbours (k-NN), Linear Regression, XG Boost, Naive Bayes, Decision Tree, Random Forests and Support Vector Machine (SVM)[6].

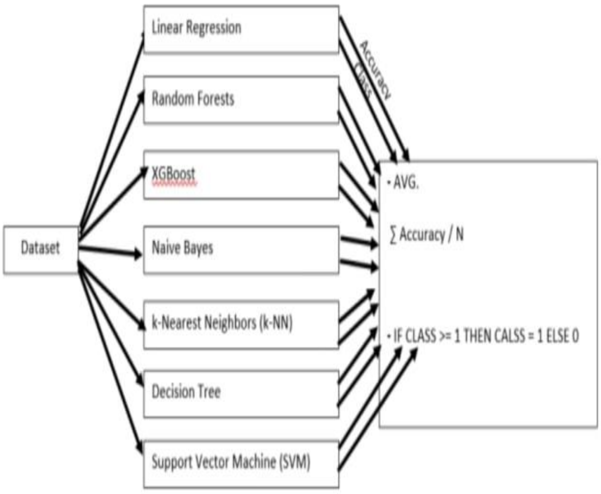


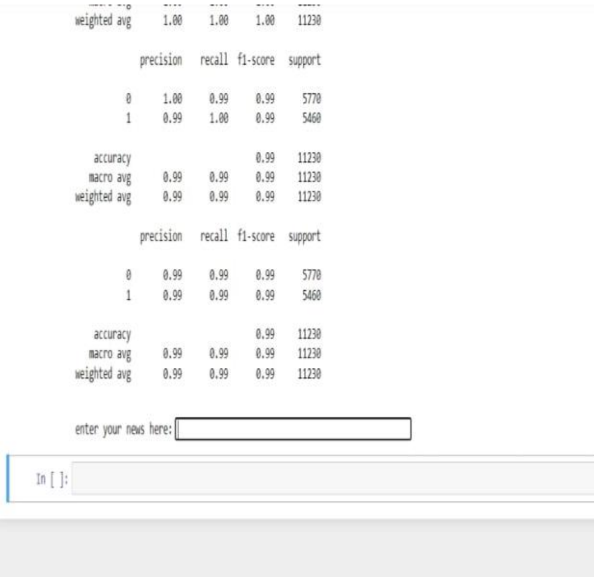
Figure 3. The Classification Algorithms

All these algorithms get as accurate as possible. Where reliable from the combination of the average of them and compare them. As shown in the figure below, the dataset is applied to different algorithms in order to detect a fake news. The accuracy of the results obtained are analysed to conclude the final result.
In this project while detecting the fake news we are taking four algorithms are logistic regression, decision tree, gradient boosting and random forest classifiers using machine learning.

VII. OUTCOME

In this project, the overall detection as a result it gives as output that is implied as outcome. In Fake news detection using classification four algorithms are logistic regression, decision tree, gradient boosting and random forest by predicting the test and importing the libraries it gives as the classification of prediction.

After implementing the four modules of four algorithms by using binary passive classification then giving some outcomes during the entire process of detection while taking the classifiers on linguistic impact on social media [4] (i.e communicating between the people in the virtual form).



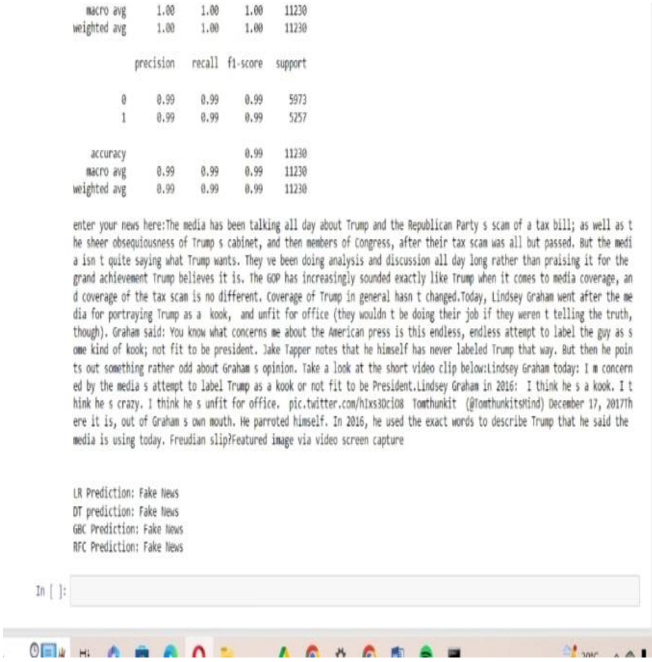
```
weighted avg       1.00      1.00      1.00     11230

              precision    recall  f1-score   support

           0       1.00      0.99      0.99      5770
           1       0.99      1.00      0.99      5460

    accuracy                           0.99     11230
   macro avg       0.99      0.99      0.99     11230
weighted avg       0.99      0.99      0.99     11230

              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5770
           1       0.99      0.99      0.99      5460

    accuracy                           0.99     11230
   macro avg       0.99      0.99      0.99     11230
weighted avg       0.99      0.99      0.99     11230

enter your news here:
```

Figure 4. Output1

As you can see the above figure, its completely given as output where you attempt the code in the python using libraries and predict the algorithm classifiers, as you can observe where its written as ‘enter your news here’ that means where the news of fake and true are kept in the file called zip file and you need to copy the link as a CSV files of data sets.

By approaching the proposed framework in the fake news detection on social media, the data set values getting the accuracy values and weighted average value by predicting the four types given classifications following Steps.

In given excel sheets of fake news and true news, you need take the text of fake news and it will be represented as a fake news in all four classifier algorithms. If you take the text of true news then it will be represented as a Not a Fake news in all four classifier algorithms. In given two diagrams are the screenshots of outputs giving the score of data sets values of fake news detection.



```
   macro avg       1.00      1.00      1.00     11230
weighted avg       1.00      1.00      1.00     11230

              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5973
           1       0.99      0.99      0.99      5257

    accuracy                           0.99     11230
   macro avg       0.99      0.99      0.99     11230
weighted avg       0.99      0.99      0.99     11230

enter your news here:The media has been talking all day about Trump and the Republican Party s scam of a tax bill; as well as t
he sheer obsequiousness of Trump s cabinet, and then members of Congress, after their tax scam was all but passed. But the medi
a isn t quite saying what Trump wants. They ve been doing analysis and discussion all day long rather than praising it for the
grand achievement Trump believes it is. The GOP has increasingly sounded exactly like Trump when it comes to media coverage, an
d coverage of the tax scan is no different. Coverage of Trump in general hasn t changed.Today, Lindsey Graham went after the me
dia for portraying Trump as a  kook,  and unfit for office (they wouldn t be doing their job if they weren t telling the truth,
though). Graham said: You know what concerns me about the American press is this endless, endless attempt to label the guy as s
ome kind of kook; not fit to be president. Jake Tapper notes that he himself has never labeled Trump that way. But then he poin
ts out something rather odd about Graham s opinion. Take a look at the short video clip below:Lindsey Graham today: I m concern
ed by the media s attempt to label Trump as a kook or not fit to be President.Lindsey Graham in 2016:  I think he s a kook. I t
hink he s crazy. I think he s unfit for office.  pic.twitter.com/hIxs3DciO8  Tomthunkit  (@TomthunkitsMind) December 17, 2017Th
ere it is, out of Graham s own mouth. He parroted himself. In 2016, he used the exact words to describe Trump that he said the
media is using today. Freudian slip?Featured image via video screen capture


LR Prediction: Fake News
DT prediction: Fake News
GBC Prediction: Fake News
RFC Prediction: Fake News
```

Figure 5. Output2

VIII. CONCLUSIONS

In this conclusion, it will represent to detect the whether the social media news is fake or not using machine learning while taking the passive aggressive algorithm classification using linguistic, content and social-context based approach in social media like Instagram, Facebook, Twitter, WhatsApp forward messages etc.

Fake news detection helps us to classify the news as fake or real. In our method, we first input the data into a trained model. We then find that the further input features are essential to achieve high prediction accuracy. We found the various linear models by using statistical, machine learning and deep learning-based methods and evaluated the performance with three metrics: accuracy and F1 score, precision score.

Due to increasing use of internet, it is now easy to spread fake news. A huge number of persons are regularly connected with internet and social media platforms. There is no any restriction while posting any news on these platforms. So some of the people takes the advantage of these platforms and start spreading fake news against the individuals or organizations.

This can destroy the repute of an individual or can affect a business. Through fake news, the opinions of the people can also be changed for a political party. There is a need for a way to detect these fake news. Machine learning classifiers are using for different purposes and these can also be used for detecting the fake news. The classifiers are first trained with a data set called training data set. After that, these classifiers can automatically detect fake news. In this systematic literature review, the supervised machine learning classifiers are discussed that requires the labeled data for training. Labeled data is not easily available that can be used for training the classifiers for detecting the fake news. In future a research can be on the use of the unsupervised machine learning classifiers for the detection of fake news.

REFERENCES

[1] P. Biyani, K. Tsioutsiouliklis, and J. Blackmer, '''8 amazing secrets for getting more clicks': Detecting clickbaits in news streams using article informality,'' in Proc. AAAI Conf. Artif. Intell., vol. 30, 2016.

[2] L. Breiman, ''Random forests,'' Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[3] T. Chen, T. He, M. Benesty, V. Khoti Lovich, Y. Tang, and H. Cho, ''Xgboost: Extreme gradient boosting,'' R Package Version vol. 1, no. 4, pp. 1–4, Aug. 2015.

[4] Y. Chen, N. J. Conroy, and V. L. Rubin, ''Misleading online content: Recognizing clickbait as 'false news,''' in Proc. 2015 ACM Workshop Multimodal Deception Detection, 2015, pp. 15–19.

[5]M.Gupta, P.Zhao, and J.Han, ''Evaluating event credibility on Twitter,'' in Proc. SIAM Int. Conf. Data Mining, Apr. 2012, pp. 153–164.

[6]T.G.Dietterich,''An experimental comparison of three methods for con- structing ensembles of decision trees: Bagging, boosting, and randomization,'' Mach. Learn., vol. 40, no. 2, pp. 139–157, 2000.

[7] R. E. Wright, Logistic Regression. Washington, DC, USA: American Psychological Association, 1995.

[8] A. L. Ginsca, A. Popescu, and M. Lupu, ''Credibility in information retrieval,'' Found. Trends Inf. Retr., vol. 9, no. 5, pp. 355–475, 2015.